Network: Infection and Inflammation: from Pathogen-induced Signatures to Therapeutic Target Genes

Project: Data Management

Gunter Hempelmann - Justus Liebig University, Giessen - gunter.hempelmann@chiru.med.uni-giessen.de

Introduction

The development of medical research networks within the framework of translational research has fostered interest in the integration of clinical and biological data in common databases [1;2]. Clinical data are required to analyse the relevance and importance of findings made in experimental research. Within the scope of the German "Nationales Genomforschungsnetz 2 (NGFN 2)", the University Hospital Giessen participates in the Infection and Inflammation Network dealing with molecular biology of infection and inflammation in different clinical pictures. Its goal is the generation of prognostic scores concerning sepsis from genomic and clinical data. The aim of this project is the definition of requirements for and the realization of a data warehouse concept dealing with data collected from clinical documentation combined with data resulting from genomic research.

Results/Project Status

In order to comply with the requirements, a concept for a data warehouse model which integrates clinical and genomic research data was chosen following the principles of incremental architected data marts [6]. The first step was focused on the data mart for clinical data. The data of the clinical documentation from intensive care units are stored electronically via a Patient Data Management System (PDMS) [3].

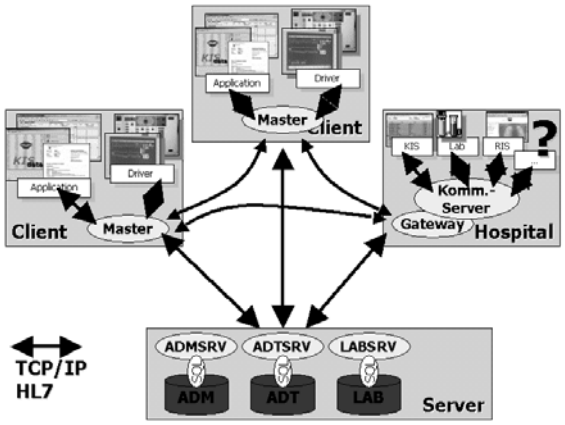

Fig 1: *The architecture of the PDMS*

The data collected by the PDMS are stored routinely in an Oracle™ database. A daily updated image of these databases is provided on a separate computer for clinical research using Oracle™ distributed computing techniques. Based on the mirrored database, the data are transformed into the data warehouse design for structural and semantic standardization. This level of the data warehouse is meant only for administrative issues. In a next step, patients are pseudonymisized and personal data are extinguished. On this pseudonymization level, user access is granted to domain specialists. A study flag assigned in an exclusive sector of the PDMS by the research physician recruits the patients to the study. For recruited patients, clinical data according to the study protocol are selected by domain specialists to populate the study data mart. On the data mart level, access is granted to researchers for data queries and data extraction.

Data derived from microarray experiments are stored in a proprietary database (PIRO-DB). The data sets are labelled with a pseudonym right from the beginning of genomic examination. Genomic data are connected with the clinical data by the trusted research physician, who is linking both pseudonyms. He is the only one who is able to re-identify a patient.
Actually, the trustee administers the pseudonyms on the administrative level. A web interface to administer pseudonyms without direct access to the database is scheduled. Furthermore, pseudonyms and remaining data are separated from each other into two different databases, according to the data security concept of the "Telematikplattform für Medizinische Forschungsnetze e.V." (TMF – national medical IT board) [7].
Scientists using the data warehouse are and will not be able to re-identify patients. They only will gain insight into the data adapted to their individual scientific problem.

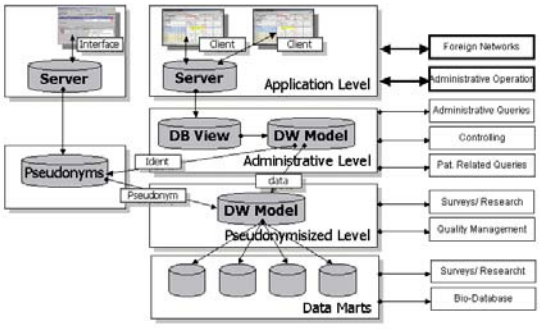

Fig 2: *Organization of the scheduled clinical data warehouse*

As we chose a non-hypothetical approach to scientific problems, the progress of the project will continuously create new scientific questions and problems; new technologies could contribute to the scientific work in the laboratories as well as to the data management. Therefore, it is a vital requirement to provide the possibility of expanding the database on a basic level. To support this dynamic process, we chose the EAV model which is defined by a flat structure of organization. Thus, it is much more easily expandable than a data model consisting of a more rigid structure, e.g. a strictly relational database. The data derived from the PDMS are already stored in a data model based on HL-7. In earlier work, existing clinical scores were successfully calculated on data from the PDMS and procedures for the score generation were validated [8].
The development of the data warehouse presented was led by the objectives performance, transparency and security. These goals were realized by the concept of a data warehouse organized in levels. Mirroring of the production database on the administrative level reduces the workload for the database of the PDMS and guarantees a complete set of data on which necessary standardizations are possible. Pseudonymization on the next level provides data

security and data privacy. Selection of patients and data according to the study protocol on the data mart level provides an efficient and comprehensive view for the researcher and allows linkage between the clinical and genomic data.
Besides projects for quality improvement and outcome studies [9], there are no known projects utilising routine clinical data from PDMSs for data warehouses. The integration of clinical data from PDMSs and genomic data for biological research broadens the fundament of data for research, although its usage obtains some obstacles. The semantic standardization at the administrative level needs ongoing refinements and should conform to semantic standards like Logical Observation Identifiers Names and Codes (LOINC), which are not implemented at the current version.
In case relevant diagnostic findings are revealed in the course of the study, a pseudonymization of the patients is more preferable than an anonymization in order to re-identify the patient. The information gained from the genomic examination could possibly influence the patient's therapy and thus affect the course of his or her disease. In this situation, it could be considered non-ethical to withhold this information from the patient.
The data marts containing the clinical data which are defined by the study protocol provide the researchers with the needed data. Actually, data are still extracted from the data marts into flat files as well as from the proprietary database for microarray research data. These flat files are transferred into tools for analysis. However, the development of web interfaces is projected in order to provide direct access to the data marts for the researchers.
Several tools exist for microarray viewing and analysis, featuring interfaces to process data formatted according to established standards like MAGE-ML [9]. For clinical data, the Clinical Data Interchange Standards Consortium (CDISC) and the Health Level 7 group (HL7) are designing data standards to enable system intraoperability. Assuming the use of analytic tools operating with standardised data, the implementation of the interfaces to exchange data in a standardised format is scheduled. These interfaces will also enable data exchange with external data sources.

Outlook

The data warehouse presented provides an efficient, transparent and secure access to the clinical data collected by the PDMS. The working basis of the data warehouse concept is data security and protection of data privacy according to valid data security laws. The level of organised data processing as well as the flexibility of an EAV model database based on the HL7 RIM allows for an integration of all resulting data, clinical data and genomic data. Since score generation and score computation based on clinical data from the PDMS using preassigned views and procedures has already been successfully realized, these concepts can now be extended to the integration of the genomic research data.

*Lit.: **1.** Altman RB et al. Challenges for Biomedical Informatics and Pharmacogenomics. Annu Rev Pharmacol Toxicol. 2002;42:113-33. **2.** Yue L et al. Pathway and ontology analysis : emerging approaches connecting transcriptome data and clinical endpoints. Bioinformatics. 2005;5(1):11-21. **3.** Michel et al. Design principles of a clinical information system for intensive care units (ICU Data). Stud Health Technol Inform. 2000;77:921-4. **6.** Hackney D. Architectures and approaches for succesful data warehouses. Oracle White Papers. 2002. **7.** Pommerening K et al. Secondary use of the electronic health record via pseudonymisation. In: Bos l et al. (eds.): Medical care Compunetics 1. Amsterdam: IOS Press. 2004:441-6. **8.** Junger et al. Automatic calculation of a modified APACHE II score using a patient data management system (PDMS). Int J Med Inform. 2002;65:145-57. **9.** Tjandra d et al. An XML message broker framework for exchange and integration of microarray data. Bioinformatics: 2003;19(14):1844-5.*